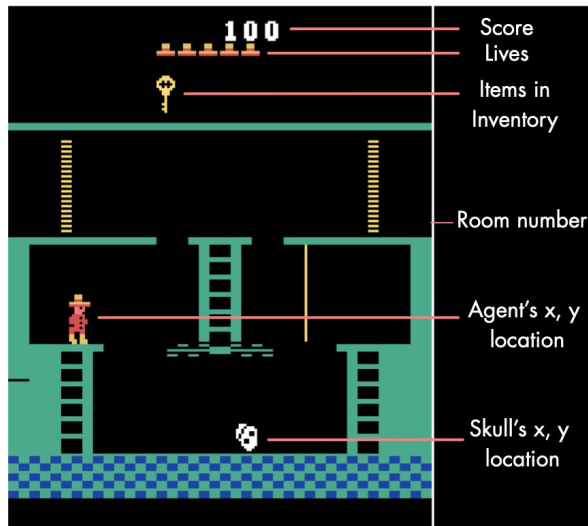


Unsupervised State Representation Learning in Atari

<https://arxiv.org/abs/1906.08226>



Ankesh Anand*, Evan Racah*, Sherjil Ozair*,

Yoshua Bengio, Marc-Alexandre Côté, R Devon Hjelm



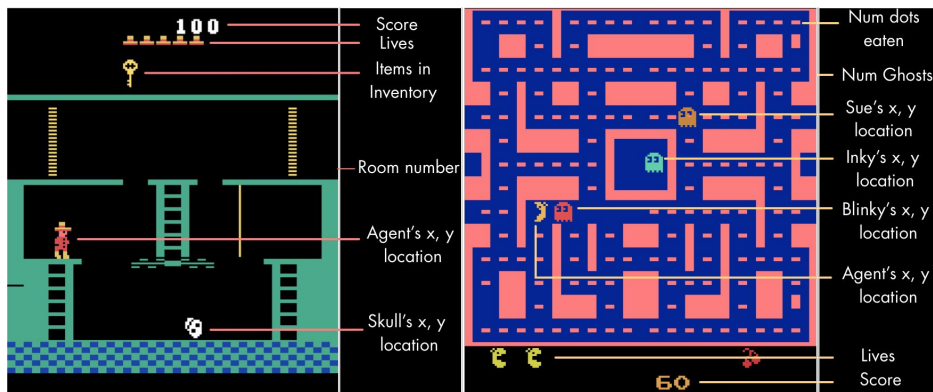
Microsoft®
Research

Key Points

- A state representation learning algorithm to learn **high-level concepts** in a scene
 - **without labels or rewards**
 - **without modelling pixels directly.**
- The **Atari Annotated RAM Interface (AARI)**: A **benchmark** to systematically evaluate state representations.

State Representation Learning

Goal: Encode high-dimensional obs. to a latent space that captures underlying generative factors of an environment




- Allow agents to learn to act in environments with **fewer interactions**.
- Effectively **transfer** knowledge across different tasks in the environment

Supervised -> Self-supervised / Unsupervised Learning

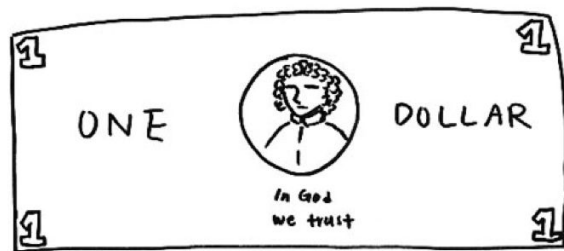
- Can't rely on direct supervision given the high dimensional nature of problems.
 - The **marginal cost** of acquiring labels in RL is much higher.
- The underlying data has a much richer **structure** than what **sparse** labels / rewards could provide.
 - Sparse signals -> sample inefficiency.
- Leads to **task-specific** policies, rather than knowledge that could be repurposed.
- **Human learning** is largely unsupervised.
 - The Scientist in the Crib: What Early Learning Tells Us About the Mind
(Alison Gopnik, Andrew N. Meltzoff and Patricia K. Kuhl, 1999)
 - The Development of Embodied Cognition: Six Lessons from Babies
(Linda Smith and Michael Gasser, 2005)

How Much Information Does the Machine Need to Predict? Y LeCun

- **"Pure" Reinforcement Learning (cherry)**
 - ▶ The machine predicts a scalar reward given once in a while.
 - ▶ **A few bits for some samples**
- **Supervised Learning (icing)**
 - ▶ The machine predicts a category or a few numbers for each input
 - ▶ Predicting human-supplied data
 - ▶ **10→10,000 bits per sample**
- **Unsupervised/Predictive Learning (cake)**
 - ▶ The machine predicts any part of its input for any observed part.
 - ▶ Predicts future frames in videos
 - ▶ **Millions of bits per sample**



Illustrative Example



From memory



One dollar bill



From reference

Representation Learning in humans doesn't seem to be operating in the pixel space.

Contrastive Unsupervised Representation Learning

- Goal: Learn an embedding function f such that:

$$\langle f(x), f(x^+) \rangle \gg \langle f(x), f(x^-) \rangle$$

- x and x^+ are similar data points [Positive].
 - x^- is a random data point (and thus presumably dissimilar to x) [Negative].
- If we use multiple negative samples, we get a lower bound on **Mutual Information**.

$$I(X, X^+) \geq \sum_{i=1}^N \log \frac{\exp f(x_i)^T f(x_i^+)}{\sum_{j=1}^N \exp f(x_i)^T f(x_j^-)} \triangleq \mathcal{I}_{NCE}(\{(x_i, x_i^+)\}_{i=1}^N)$$

- To maximize MI, we can compute gradients of this lower bound on MI w.r.t a parametric encoder f_θ .

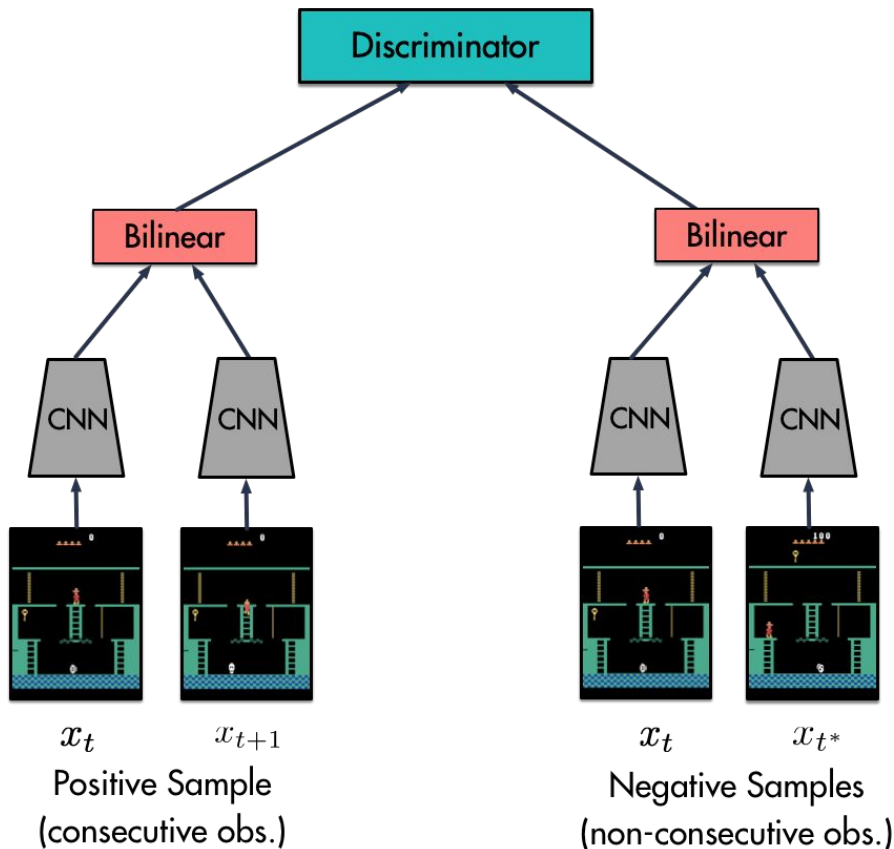
Arora et. al (CURL) ICML '19
Poole et. al (MI Bounds) ICML '19
Hjelm et. al (Deep InfoMax) ICLR '19
Van Den Oord et. al (CPC) 2018

Nature of RL environments

- **Temporal** Structure in Data.
 - Not i.i.d sampled.
- **Local consistency**. Objects don't move drastically over single time-steps.
- Prior work has argued that **without** auxiliary variables (such as time), recovering underlying latent variables is generally not possible.
 - Hyvarinen et. al (2014)
 - Locatello et. al (2019)

Can we exploit the inherent temporal structure to learn representations?

The contrastive task



- x_{t^*} is randomly sampled from the episode.
- In practice, we use multiple negative samples.
- Standard CNN architecture from Minh et. al (2014) [DQN]

$$I(f(x_t), f(x_{t+1}))$$

Temporal InfoMax

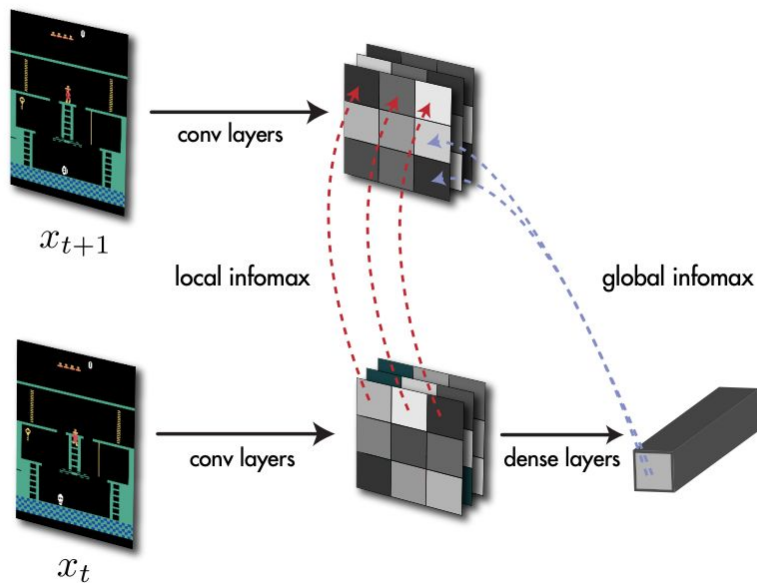
Temporal InfoMax is not enough

- The encoder can “cheat” and focus on just one factor of variation that’s easy to predict (like the clock).
- We incorporate a **spatial prior** to incentivize the encoder to focus on **all** factors of variation.



Spatio-Temporal DeepInfoMax (ST-DIM)

- Maximize the temporal MI **spatially** across each local feature map.
- Each feature map has a receptive field corresponding to $1/16^{th}$ the size of the full of the observation.

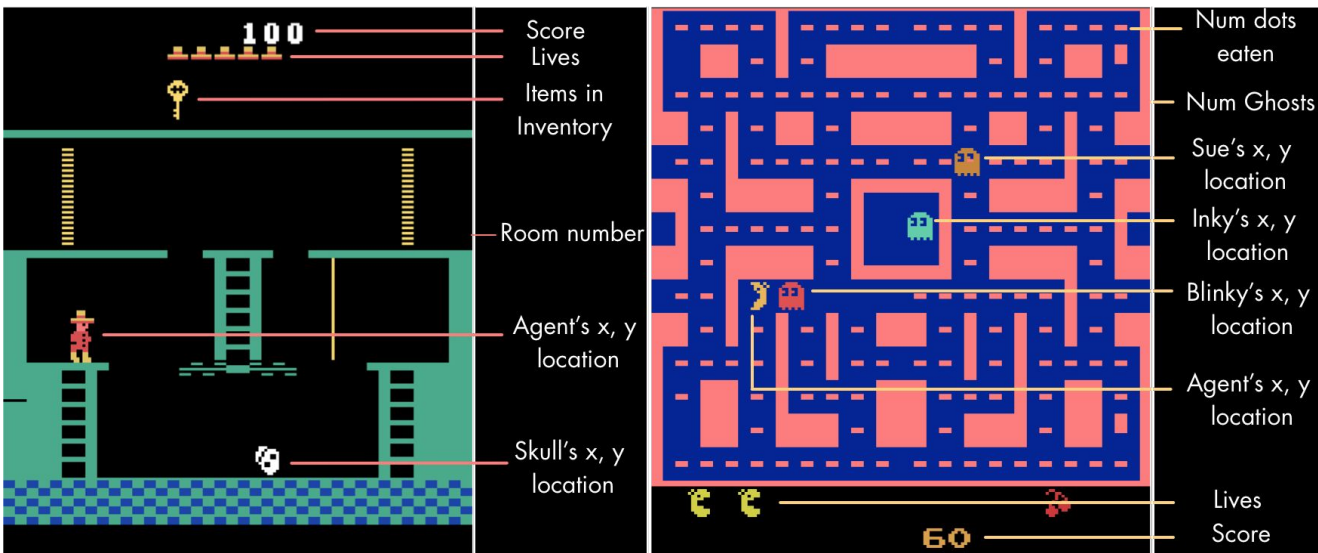


Evaluating Representations

- Evaluating representations is hard.
- Performance on a single downstream task (e.g. control with a single reward function, next-frame prediction)
 - Might not measure all useful things a representation should capture
- More principled approach: measure ability of a representation at **capturing all high-level factors.**

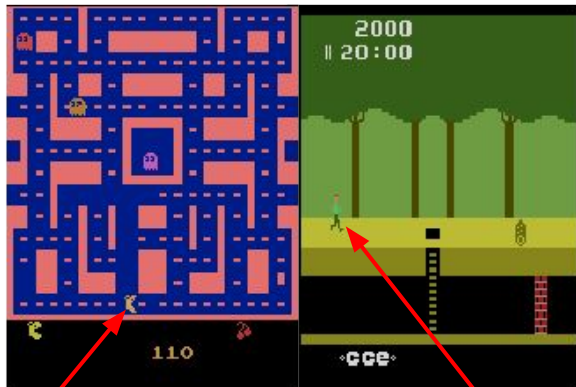
Atari Annotated Ram Interface (AARI)

- Interface for evaluating state representations in Atari games
- Gym wrapper exposes 308 total semantic labels, for 22 Atari games
- Evaluation using Linear Probing.



Categorization of State Variables

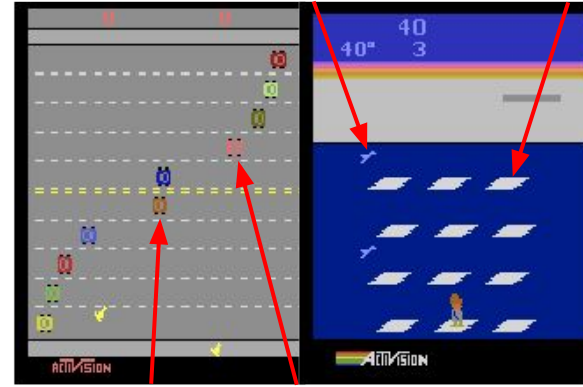
Agent Localization



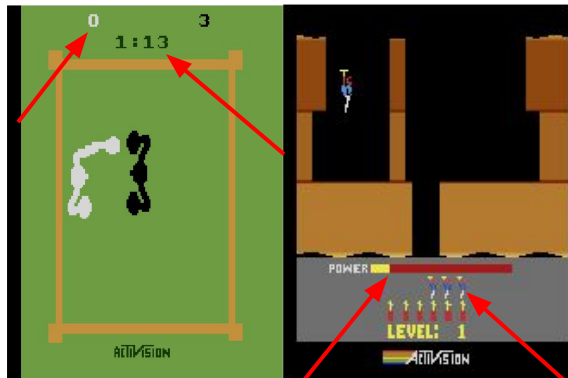
Small Object Localization



Other Localization



Score/Clock/Lives/Display



Miscellaneous



State Variable Breakdown

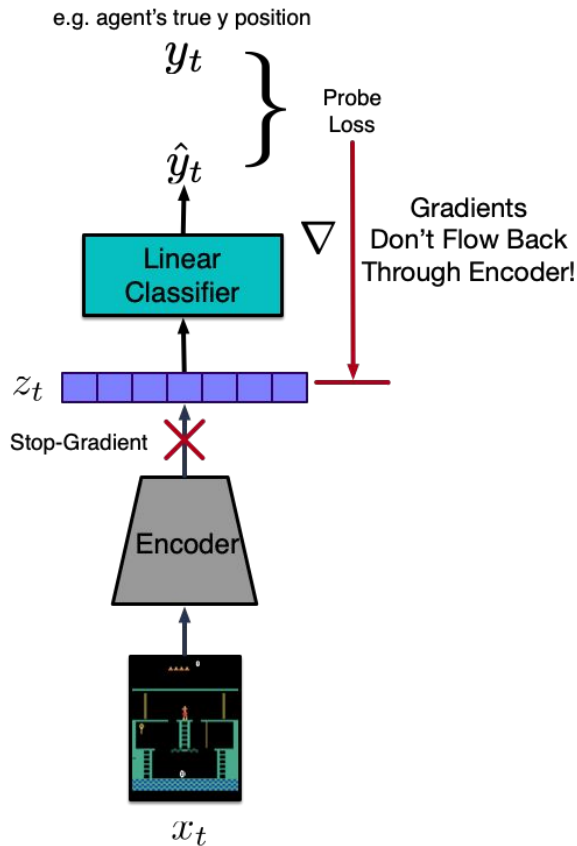
- 22 Total Games
- 308 Total State Variables

GAME	AGENT LOCAL.	SMALL OBJECT LOCAL.	OTHER LOCAL.	SCORE/CLOCK		OVERALL
				LIVES DISPLAY	MISC	
ASTEROIDS	2	4	30	3	3	41
BERZERK	2	4	19	4	5	34
BOWLING	2	2	0	2	10	16
BOXING	2	0	2	3	0	7
BREAKOUT	1	2	0	1	31	35
DEMONATTACK	1	1	6	1	1	10
FREEWAY	1	0	10	1	0	12
FROSTBITE	2	0	9	4	2	17
HERO	2	0	0	3	3	8
MONTEZUMAREVENGE	2	0	4	4	5	15
MSPACMAN	2	0	10	2	3	17
PITFALL	2	0	3	0	0	5
PONG	1	2	1	2	0	6
PRIVATEEYE	2	0	2	4	2	10
QBERT	3	0	2	0	0	5
RIVERRAID	1	2	0	2	0	5
SEAQUEST	2	1	8	4	3	18
SPACEINVADERS	1	1	2	2	1	7
TENNIS	2	2	2	2	0	8
VENTURE	2	0	12	3	1	18
VIDEOPINBALL	2	2	0	2	0	6
YARSREVENGE	2	4	2	0	0	8
TOTAL	39	27	124	49	70	308

Evaluation Using Probing

- We focus on measuring “explicitness”
 - to what extent true state can be recovered from learned representation using linear transformation

1. Freeze encoder's weights,
2. Train a **linear classifier** on top of each representation



Training Details

- Unsupervised training for 100K frames before probing
- Probe train-val-test -> 50K frames
- Two data collection modes:
 - Random policy
 - Expert PPO policy
- Probing-Prune state variables that don't vary in value very much (entropy pruning)
- Different linear classifier for each state variable
- F1 Score to account for label imbalance

Baselines

- Majority Classifier
- Random-CNN
- VAE
- Pixel-Prediction
- CPC (Contrastive Predictive Coding)
- Fully Supervised (Upper Bound)

Results

Table 2: Probe F1 scores averaged across categories for each game (data collected by random agents)

GAME	MAJ-CLF	RANDOM-CNN	VAE	PIXEL-PRED	CPC	ST-DIM	SUPERVISED
ASTEROIDS	0.28	0.34	0.36	0.34	0.42	0.49	0.52
BERZERK	0.18	0.43	0.45	0.55	0.56	0.53	0.68
BOWLING	0.33	0.48	0.50	0.81	0.90	0.96	0.95
BOXING	0.01	0.19	0.20	0.44	0.29	0.58	0.83
BREAKOUT	0.17	0.51	0.57	0.70	0.74	0.88	0.94
DEMONATTACK	0.16	0.26	0.26	0.32	0.57	0.69	0.83
FREEWAY	0.01	0.50	0.01	0.81	0.47	0.81	0.98
FROSTBITE	0.08	0.57	0.51	0.72	0.76	0.75	0.85
HERO	0.22	0.75	0.69	0.74	0.90	0.93	0.98
MONTEZUMAREVENGE	0.08	0.68	0.38	0.74	0.75	0.78	0.87
MSPACMAN	0.10	0.49	0.56	0.74	0.65	0.72	0.87
PITFALL	0.07	0.34	0.35	0.44	0.46	0.60	0.83
PONG	0.10	0.17	0.09	0.70	0.71	0.81	0.87
PRIVATEEYE	0.23	0.70	0.71	0.83	0.81	0.91	0.97
QBERT	0.29	0.49	0.49	0.52	0.65	0.73	0.76
RIVERRAID	0.04	0.34	0.26	0.41	0.40	0.36	0.57
SEAQUEST	0.29	0.57	0.56	0.62	0.66	0.67	0.85
SPACEINVADERS	0.14	0.41	0.52	0.57	0.54	0.57	0.75
TENNIS	0.09	0.41	0.29	0.57	0.60	0.60	0.81
VENTURE	0.09	0.36	0.38	0.46	0.51	0.58	0.68
VIDEOPINBALL	0.09	0.37	0.45	0.57	0.58	0.61	0.82
YARSREVENGE	0.01	0.22	0.08	0.19	0.39	0.42	0.74
MEAN	0.14	0.44	0.40	0.58	0.61	0.68	0.82

Categorical Breakdown

- ST-DIM
 - excels at capturing **small-objects** (tough for reconstruction based methods).
 - **robust** to easy-to-exploit features.

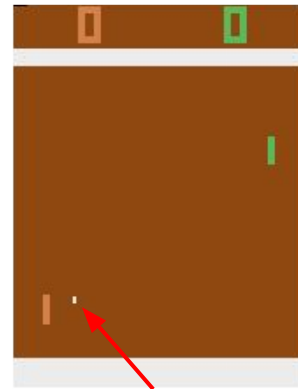


Table 3: Probe F1 scores for different methods averaged across all games for each category (data collected by random agents)

CATEGORY	MAJ-CLF	RANDOM			PIXEL-PRED	CPC	ST-DIM	SUPERVISED
		CNN	VAE					
SMALL LOC.	0.14	0.19	0.18		0.31	0.42	0.51	0.66
AGENT LOC.	0.12	0.31	0.32		0.48	0.43	0.58	0.81
OTHER LOC.	0.14	0.50	0.39		0.61	0.66	0.69	0.80
SCORE/CLOCK/LIVES/DISPLAY	0.13	0.58	0.54		0.76	0.83	0.87	0.91
MISC.	0.26	0.59	0.63		0.70	0.71	0.75	0.83

Easy-to-Exploit Features

- Contrastive losses can fail to capture all salient factors
- Especially if one factor is very easy/predictable
 - E.g. clock in Boxing
- Good litmus test: can the representation capture more than just the clock?

Table 4: Breakdown of F1 Scores for every state variable in Boxing for ST-DIM, CPC, and Global-T-DIM, an ablation of ST-DIM that removes the spatial contrastive constraint, for the game Boxing

METHOD	VAE	PIXEL-PRED	CPC	GLOBAL-T-DIM	ST-DIM
CLOCK	0.03	0.27	0.79	0.81	0.92
ENEMY_SCORE	0.19	0.58	0.59	0.74	0.70
ENEMY_X	0.32	0.49	0.15	0.17	0.51
ENEMY_Y	0.22	0.42	0.04	0.16	0.38
PLAYER_SCORE	0.08	0.32	0.56	0.45	0.88
PLAYER_X	0.33	0.54	0.19	0.13	0.56
PLAYER_Y	0.16	0.43	0.04	0.14	0.37



Future Directions

- Learning “Abstract” **world models**.
- **Sample efficient** downstream RL models.
- Do contrastive representations **generalize** better?
- What can we get with **large scale unsupervised training**?